PPFOLD PLUG-IN 2.5

USER MANUAL

Zsuzsanna Sükösd (zs@mb.au.dk)

7/1/2011

This document describes the installation and use of the PPfold plug-in in the CLC Workbenches.

PPfold Plug-in User Manual

Contents

Introduction

Welcome to PPfold -- a new version of the popular pfold program for the prediction of RNA secondary structure. PPfold is available as a standalone program as well as a plug-in to the CLC Workbenches.

1. Contact information

If you have questions or comments regarding the program, you are welcome to contact me:

Zsuzsanna Sükösd
zs@mb.au.dk

Department of Molecular Biology
Aarhus University
Aarhus C, DK-8000

2. History

PPfold has been developed in collaboration between Aarhus University, CLC bio and IT University of Copenhagen, funded by the Danish Agency for Science, Technology, and Innovation under the project "PC Mini-Grids for Prediction of Viral RNA Structure and Evolution", #09-061856.

Contributions from CLC bio by: Bjarne Knudsen, Morten Værum, Alex Andersen, Mikkel Nygaard Ravn

3. Copyright notice, licensing and source code

PPfold and the PPfold plug-in interface: This software is distributed under the “Modified BSD license”, under the following terms:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Source code for the PPfold core package is available on the PPfold website:

http://www.daimi.au.dk/~compbio/pfold/downloads.html

Source code for the plug-in interface is available on request.

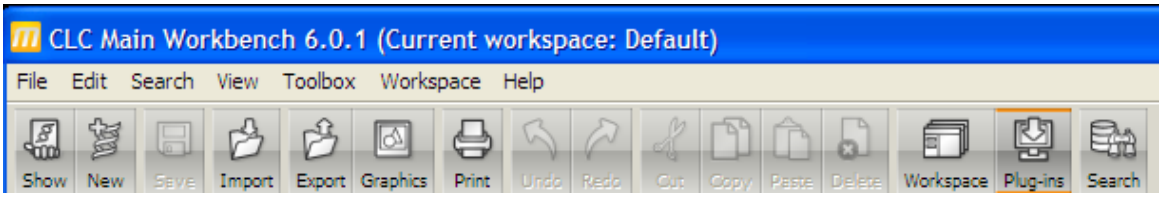
If you have used PPfold in your work and found it helpful, please cite:

Z. Sükösd, B. Knudsen, M. Værum, J. Kjems, E.S.Andersen. Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars BMC Bioinformatics 12:103, 2011

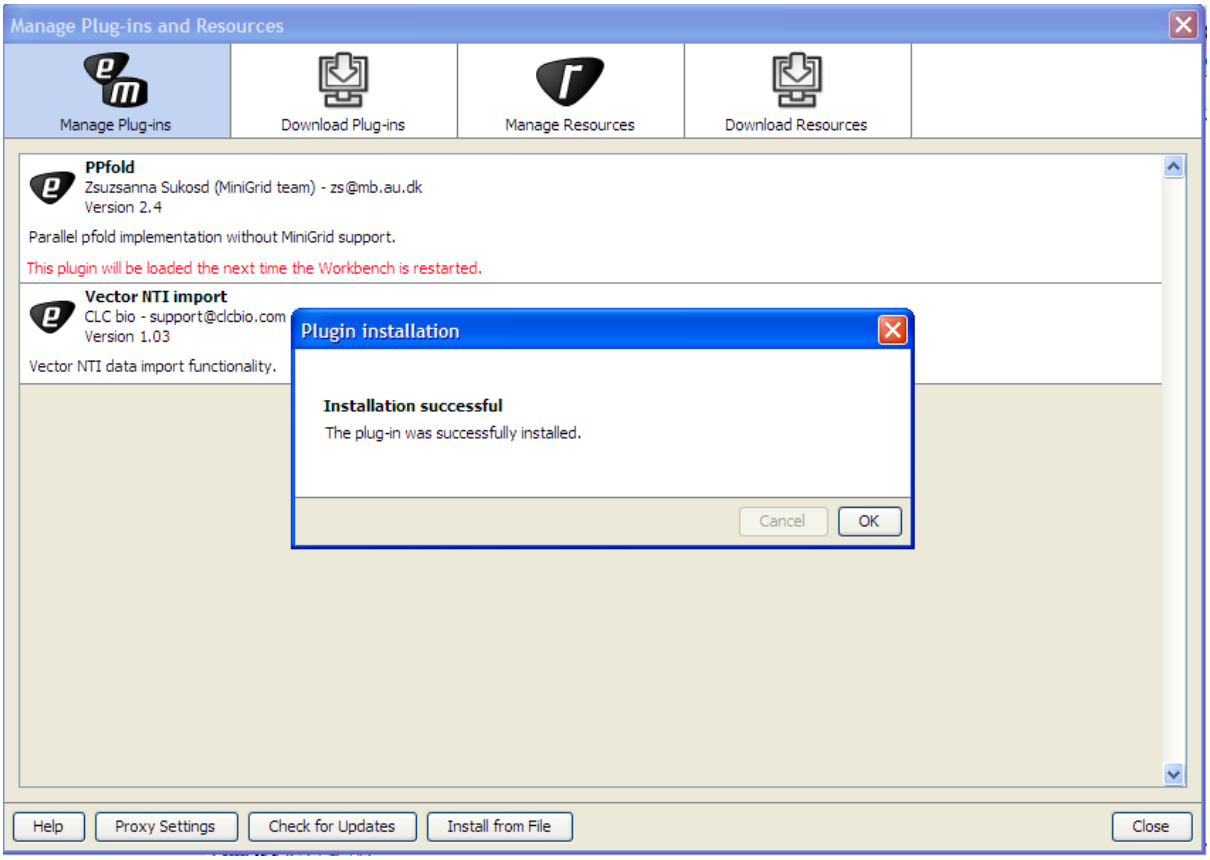
Installing the PPfold plug-in

To install the plug-in from a file:

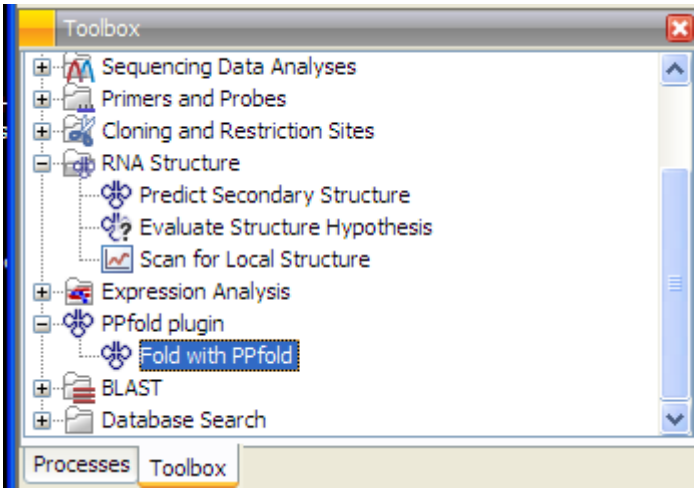
1. Download the plug-in file (.cpa), for example from the PPfold website.
2. In the CLC Workbench, click on "Plug-Ins" in the toolbar:



3. Choose "Install from File".
4. Select the plug-in file (.cpa) and click on "Install"
5. The plug-in will install automatically. You will be required to restart the Workbench to be able to use the plug-in.



6. After installation, a new command group will appear in the Toolbox, containing the "Fold with PPfold" command.



Overview: purpose, input, output and exports

1. Purpose

PPfold is a program to predict the consensus secondary structure of RNA alignments. It is a multithreaded and improved version of the popular pfold program. The PPfold plug-in enables the use of PPfold inside the CLC Workbenches.

2. Input objects

PPfold can be executed on a number of different objects:

1. **A single nucleotide sequence, or a number of nucleotide sequences**
   PPfold will fold and annotate each nucleotide sequence separately.

2. **A single nucleotide sequence list, or a number of nucleotide sequence lists**
   PPfold will fold and annotate each sequence in each sequence list separately.

3. **Any combination of 1-2**

4. **A single alignment, or a number of alignments**
   PPfold will fold and annotate each alignment separately.

5. **One alignment and one corresponding phylogenetic tree**
   PPfold will fold the alignment on the basis of optimized branch lengths in the input tree. NB: The tree must match the alignment: the names of the leaves in the tree must be in one-to-one correspondence to the names of the sequences in the alignment. It is furthermore a requirement that all branches have a length.

3. Output

PPfold modifies the input objects with annotations and provides the option of creating new objects.

1. Each individual sequence in the input is annotated with
   a. A secondary structure
   b. The probability of the secondary structure at each position

2. A dotplot of basepairing probabilities can be created
3. The maximum likelihood estimate (MLE) tree can be drawn

These can then be manipulated further in the CLC Workbenches.

4. Exports

The output of PPfold can be exported using the built-in export functions in the CLC Workbenches. In addition to these, the PPfold plug-in comes bundled with a number of custom export functions:

1. Alignments can be exported as:

   a. **Connectivity Table (.ct) format**: PPfold will attempt to identify the consensus structure in the alignment and export it in .ct format. The length of this .ct file will correspond to the length of the alignment. (If the structures of the sequences are not consistent with a consensus structure, you can potentially get strange outcomes.)

   b. **SARSE-compatible sequence (.seq) format**:

```
pairingmask ((((((...((.
gca_bovine  AGCCCUGuggUGa
gca_chicken GACUCUGuagUGa
gca_mouse   GGUCUUAaggUGa
gca_rat     AGCCUUAaggUGa
```

The first line contains the consensus pairing mask. This is followed by each sequence; basepaired nucleotides are marked with uppercase letters, single-stranded nucleotides are lowercase letters. This file is designed for import into the SARSE program.

c. **Long sequence (.lseq) format:** contains the consensus pairingmask as well as a derived structure for each sequence in dot-bracket form. This file is designed to ease the extraction of individual sequences and structures from the dataset.

d. **Position reliability (.st) format:** pfold-style export format, containing the structure and reliability score for each position for the alignment.

2. Sequences can be exported in
   a. **Position reliability (.st) format:** pfold-stype export format, containing the structure and reliability score for each position of the sequence.

3. Dotplots can be exported in
   a. **Tabbed matrix (.bp) format:** the dotplot is interpreted as basepairing probability scores generated by PPfold. The output is a tabbed matrix containing the scores at each position[1].

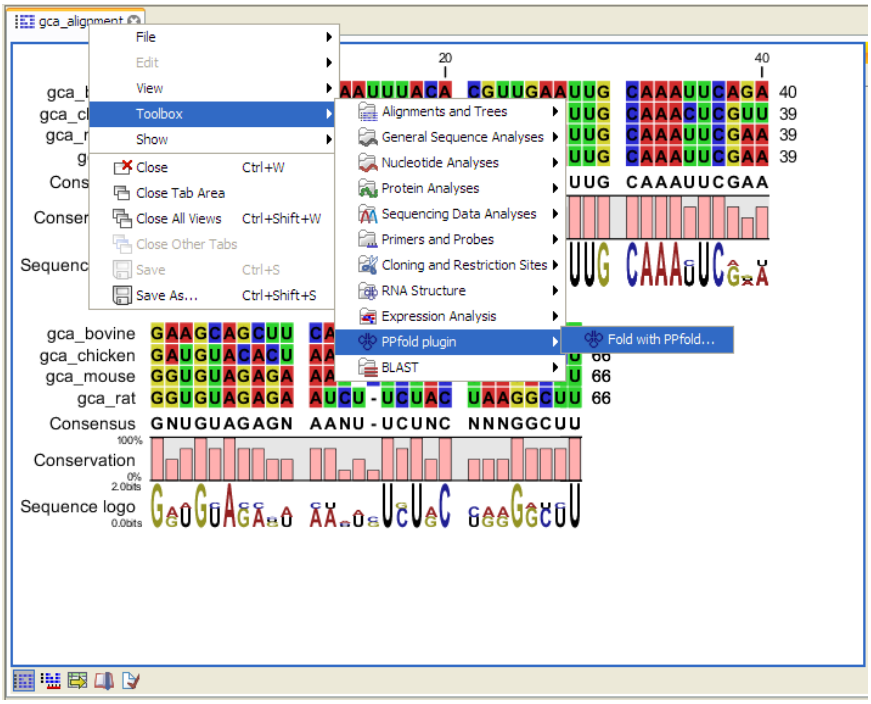
Note that all structure export functions require that the sequences are annotated with precisely one secondary structure.

[1] NOTE: If you would like to use basepairing probabilities for further calculations, I recommend using the standalone version of PPfold, as there is a significant loss of precision in creating and exporting the dotplot.

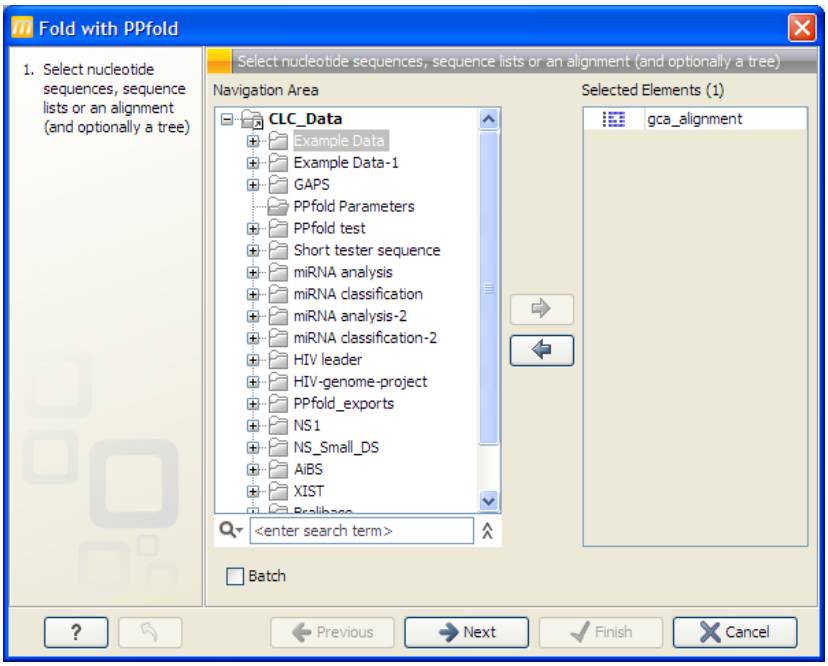
Tutorial: Folding an alignment

This tutorial will demonstrate the folding of an alignment using the PPfold plug-in in the CLC Workbenches.

1. Right-click the input alignment and select "Fold with PPfold" in the Toolbox.



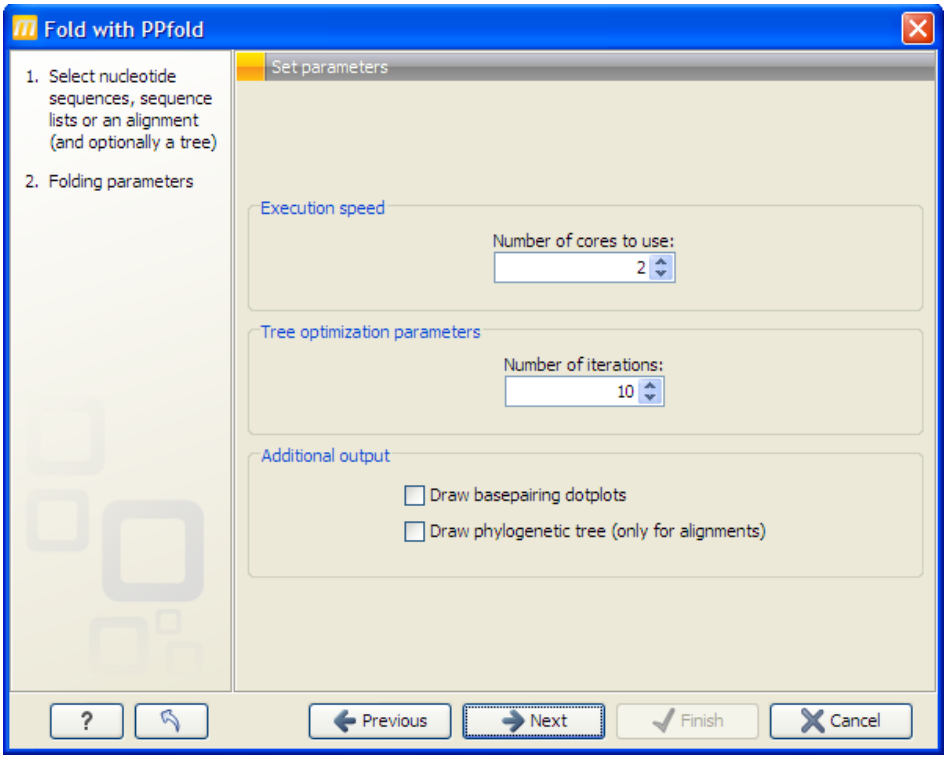
2. Click Next (unless you wish to add more input objects)



3. Choose the number of cores you wish to use in the calculations. More cores will mean faster results, but your computer will run other applications slower meanwhile. Fewer cores mean slower results, but you will be able to use your computer more effectively meanwhile.

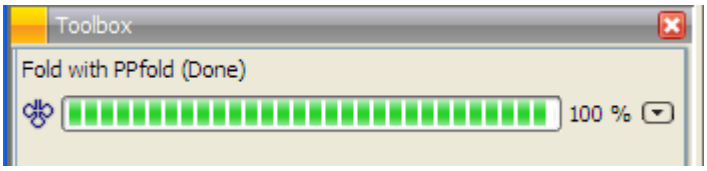
   If you have selected an alignment or alignments without a phylogenetic tree, PPfold will offer the option to select the maximum number of iterations to use in the optimization of the branch lengths of the tree. PPfold stops adjusting the branch lengths either on convergence, or when the maximum number of iterations is exceeded. In many cases, the branch lengths will converge in fewer than 10 iterations. If convergence is not obtained within 10 iterations, the tree is likely to be good enough anyway. However, increasing the number of iterations might make the tree more accurate. A higher number of iterations will mean a longer execution time, depending on the length of your alignment and the number of sequences in it.

   Select also if you wish to display basepairing dotplots, and in the case of an alignment or alignments without a phylogenetic tree, whether you wish to display the maximum likelihood estimate tree generated by PPfold. PPfold will create the data for both the basepairing dotplots and the phylogenetic tree for its own use, no matter what you choose (so it will not run faster if you choose not to display them).

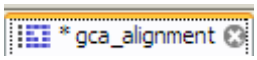


When done, click Next.

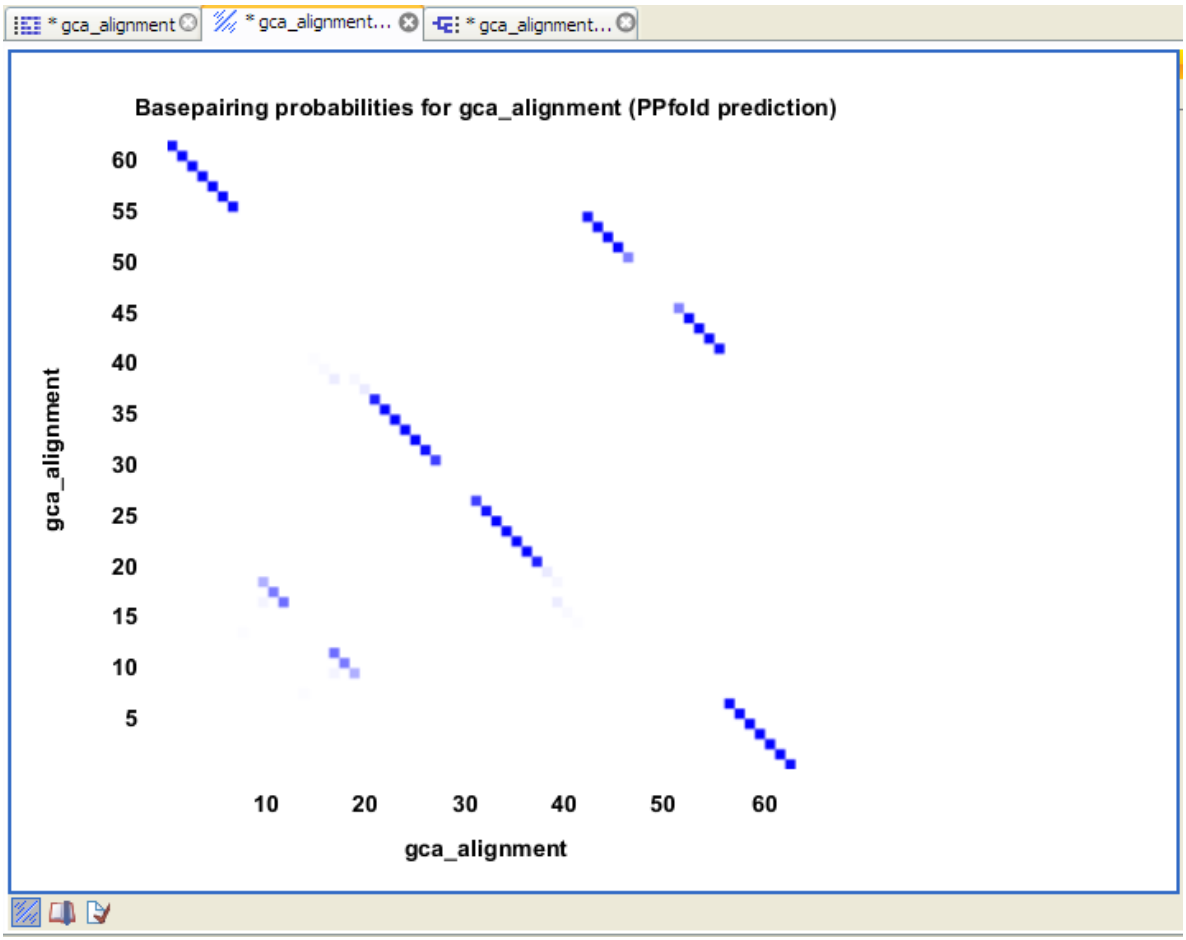
4. Select what to do with the results (Open or Save), and whether you want a log of the process.
5. Click “Finish”. PPfold will then fold the alignment. The progress of the algorithm is shown in the progress bar, which you can also use to cancel the execution.

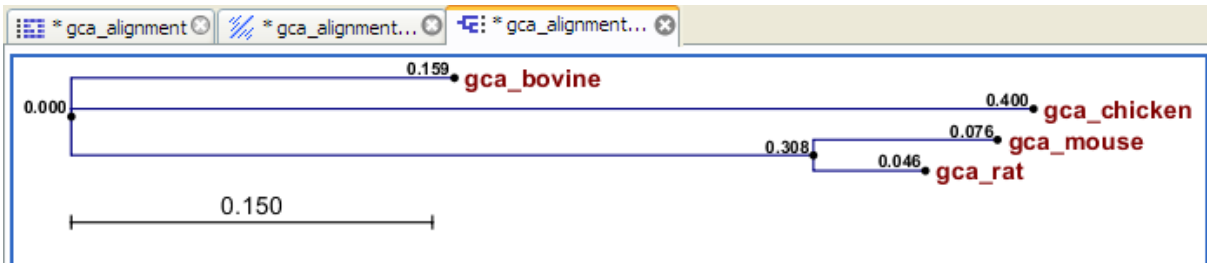


6. When the folding is finished, the alignment will be annotated with a structure. This is shown by a * next to the name of the alignment.



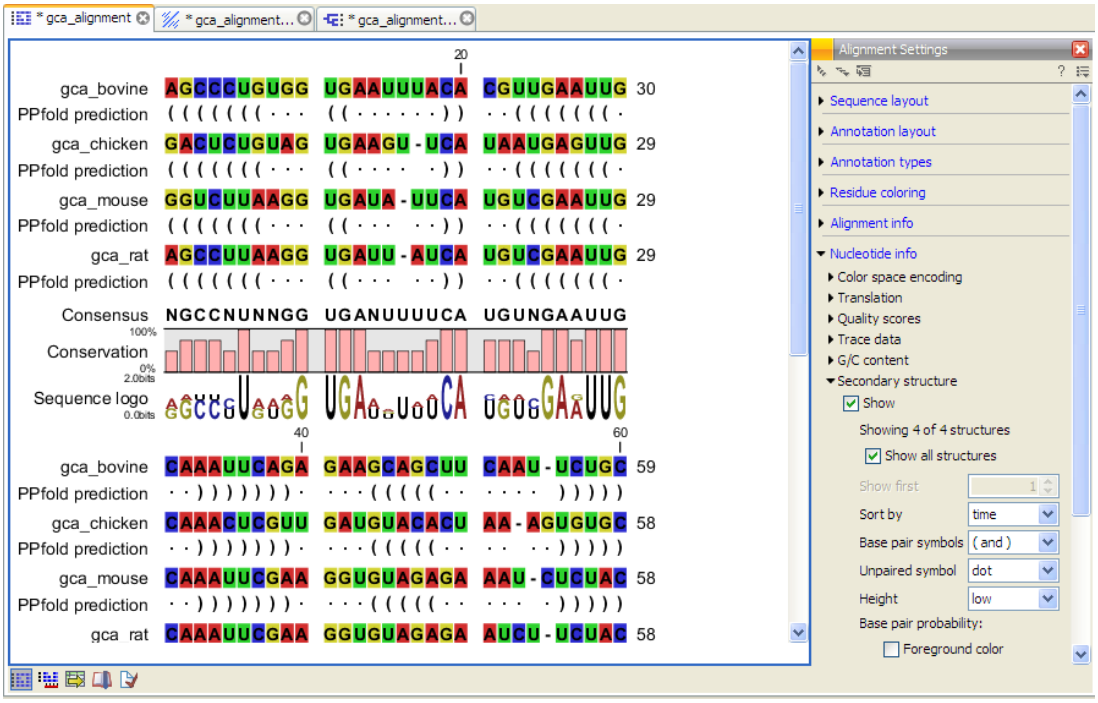
Other objects, such as a tree and a dotplot, will also have been created, if you selected them in the wizard.



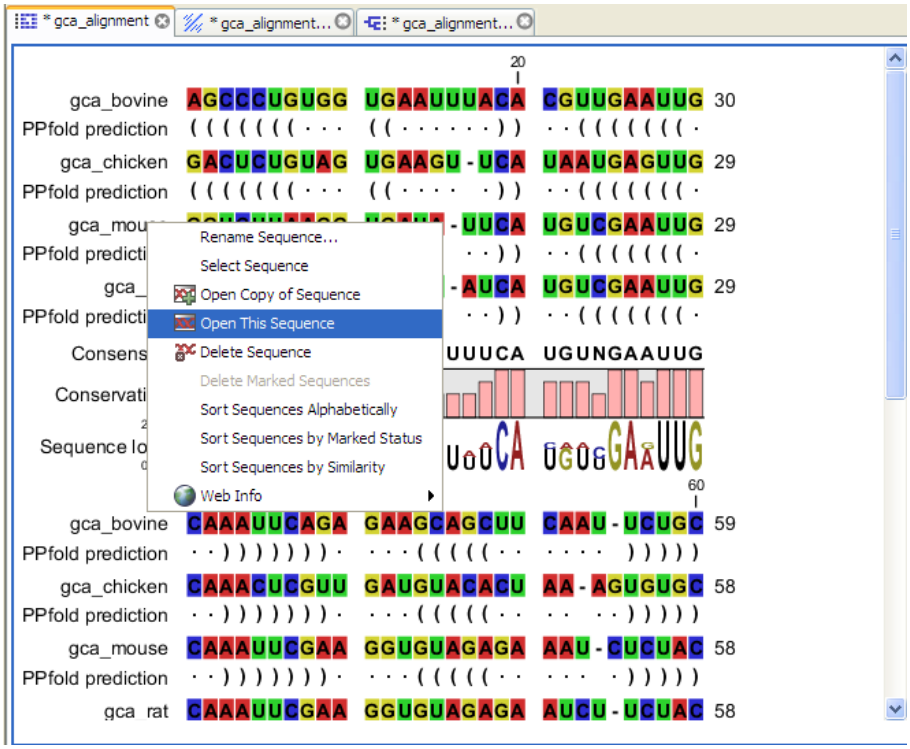


7. Select the alignment tab to view the annotated alignment.

8. To show the structure under each sequence, select: Alignment Settings -> Nucleotide info -> Secondary structure -> Show.

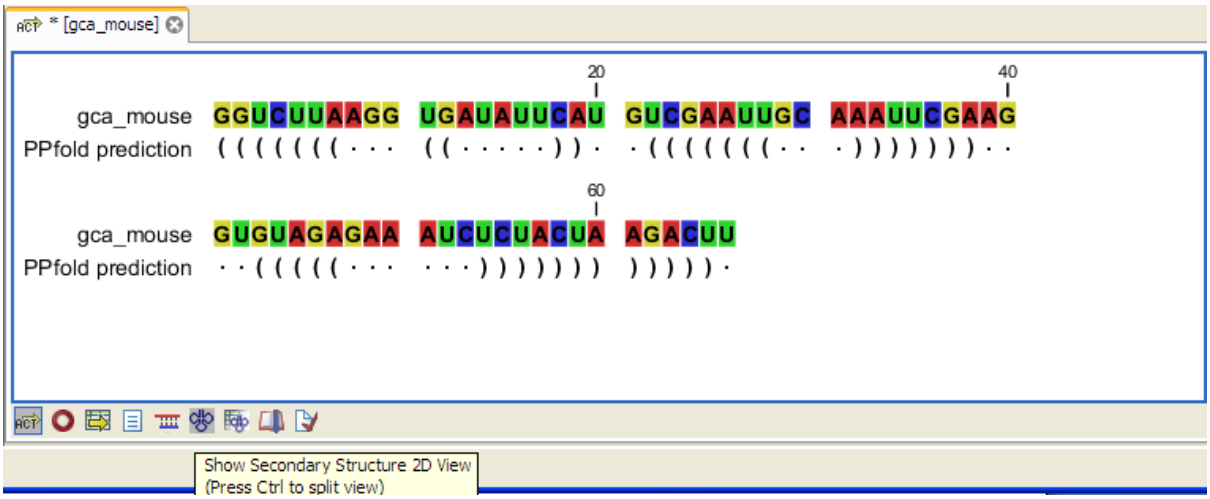


9. To display a drawing of the secondary structure, you must open a sequence first. Right-click the name of a sequence and select "Open this sequence". (You can also "Open a copy of this sequence")

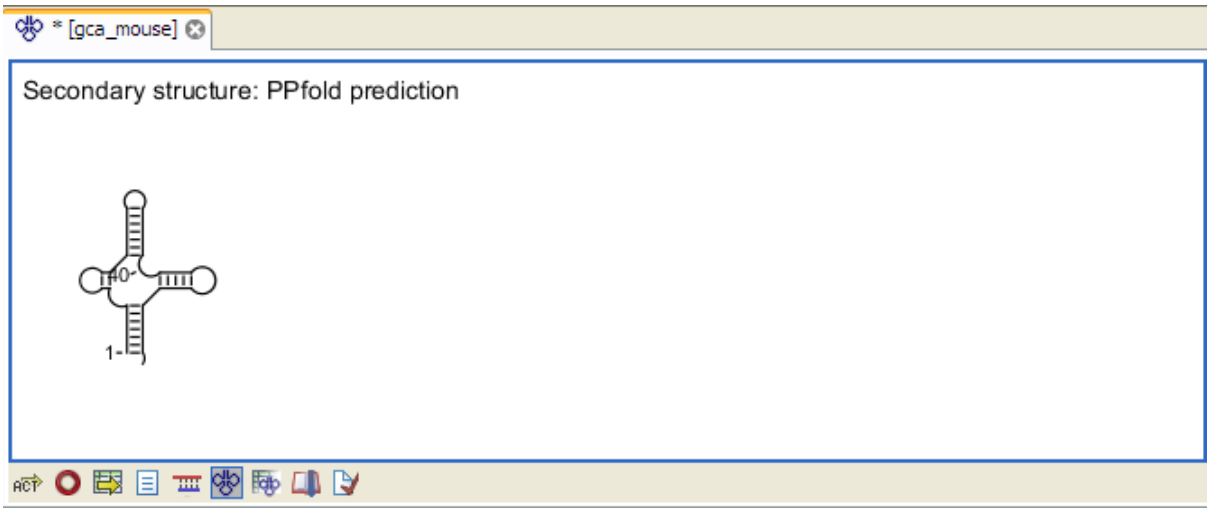


The annotated sequence will then appear.

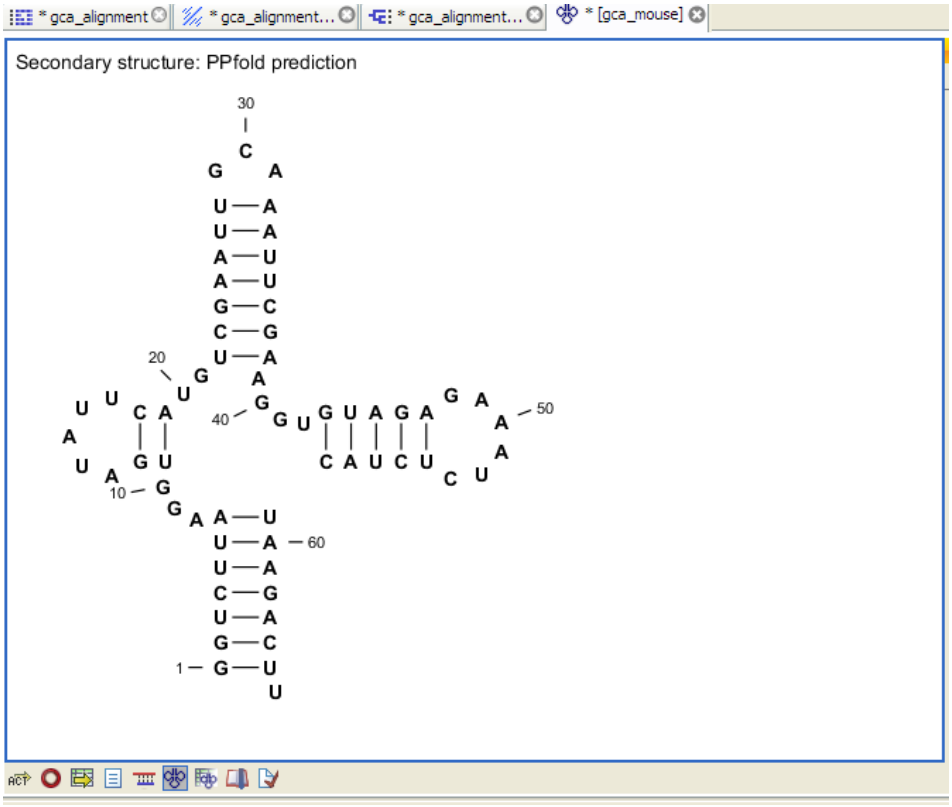
10. Click on the Secondary structure button to display the secondary structure.



The secondary structure will display:

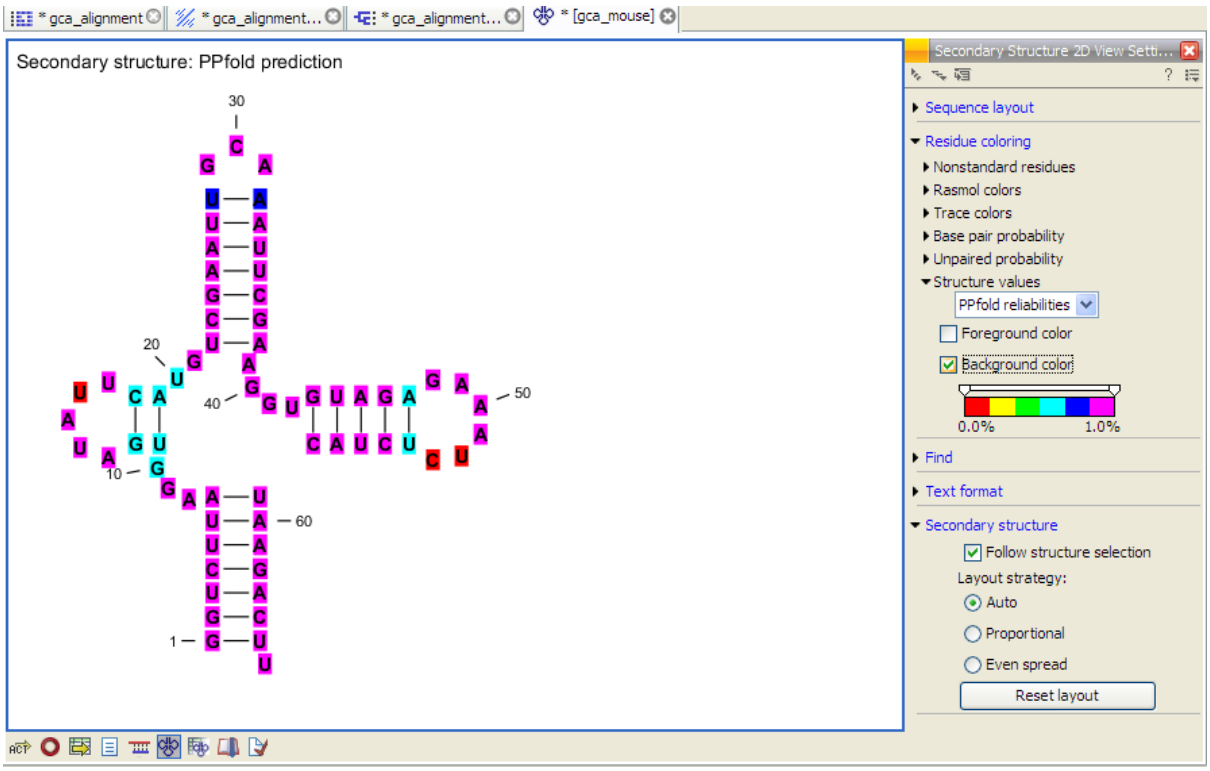


11. Zoom in to the structure to show the individual nucleotides.

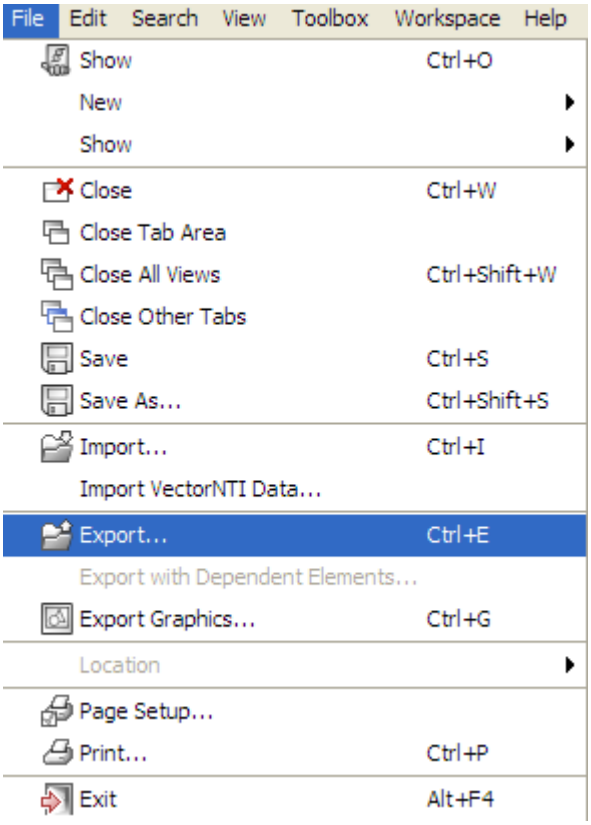


The structure drawing can be adjusted by panning structural elements.

12. To display structure reliability values as colors, choose Residue coloring -> Structure values -> PPfold reliabilities and select foreground and/or background colors.



13. To export the alignment, select the alignment and click on File -> Export.



14. Choose the desired export format from the drop-down menu and save the file.

Additional notes

The standalone version of PPfold has additional options for the advanced user, including the possibility to adjust distribution parameters or the use of an alternative parameter file. The standalone version can be downloaded from the PPfold website.